ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 1, January 2025

A Transfer Learning Framework for Biomedical Knowledge Extraction: Enhancing Accuracy, Precision, and Scalability

[1] Kasakani Yaswanth Kumar, [2] Dr. Suneetha Eluri, M.Tech, Ph.D

[1] [2] CSE, UCEK JNTUK, JNTU KAKINADA, Kakinada, Andhra Pradesh, India
Corresponding Author Email: [1] yeshupawan2001@gmailcom, [2] suneethaeluri83@jntucek.ac.in

Abstract*— In healthcare and biomedical research, enhancing the understanding of medical texts through Named Entity Recognition (NER) is vital for identifying key terms such as diseases, drugs, proteins, and genes. Transfer learning models like BERT have been instrumental in improving the extraction of domain- specific information. NER is critical for supporting medical decision-making and research. However, challenges persist, such as the absence of standardized datasets, limited computational resources, and the inherent complexity of medical data. Advanced techniques like Bidirectional Long Short-Term Memory (BiLSTM), Conditional Random Fields (CRF), and Multi-Task Learning (MTL) are employed to improve NER performance. Preprocessing involves tokenization, annotation, and embedding, while post-processing refines model predictions. The BINER model has demonstrated superior F1-scores for disease detection and has outperformed BioBERT in recognizing proteins and genes, showcasing advancements in biomedical NER applications.

Keywords*— Bidirectional Long Short Term Memory, Natural Language Processing, Named Entity Recognition, Multi-Task Learning, Biomedical Text, Gated Recurrent Unit, Convolutional Neural Network, Stochastic Gradient Descent.

I. INTRODUCTION

1.1 BioNER

Biomedical Named Entity Recognition (BioNER) plays a vital role in biomedical text mining, concentrating on the identification and classification of entities such as genes, proteins, diseases, and drugs within texts. By utilizing machine learning techniques, particularly deep learning models, BioNER systems extract essential information from biomedical documents, facilitating tasks like literature mining, clinical text processing, and database organization.

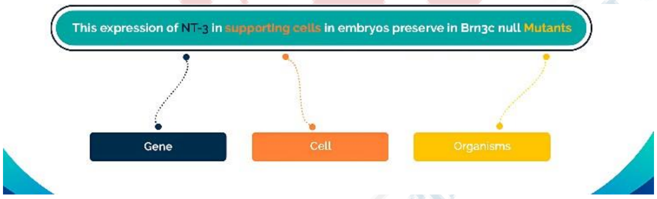

This expression of NT-3 in supporting cells in embryos preserve in Brn3c null Mutants
Gene
Cell
Organisms

Figure 1. Visualization of Named Entity Recognition in Biomedical Text

1.2 Applications of BioNER

BioNER is applied in several important areas. In literature mining, it helps researchers identify critical biomedical entities in scientific papers, making it easier to extract pertinent information. In clinical text analysis, BioNER extracts vital medical details from patient records, aiding tasks like disease tracking and adverse effect detection. For database curation, BioNER streamlines the organization and discovery of new knowledge in biomedical databases by labeling entities in documents, thereby improving the efficiency and accuracy of database management.

1.3 Challenges in BioNER

BioNER faces numerous challenges. The deep learning models it relies on demand substantial computational resources, making deployment costly. The complexity of biomedical terminology and jargon further complicates entity recognition, with varying terms across different contexts and languages leading to inconsistencies. Annotating biomedical texts is both time-consuming and expensive, limiting the availability of datasets. Biomedical entities are complex, which makes their segmentation and recognition challenging. dEnsuring that models trained on one dataset generalize well to others is challenging due to data heterogeneity. Balancing performance with computational cost while choosing the right neural network architecture and techniques is another significant hurdle.

1.4 Approaches in BioNER

To tackle these challenges, several approaches are used in BioNER. Pre-trained models like BioBERT and SciBERT are fine-tuned for specific tasks, adapting general linguistic knowledge is adapted for domain-specific applications through advanced deep learning techniques such as BiLSTM, CRF, CNN, and multi-task learning, which enhance the performance of Named Entity Recognition (NER). Word and character-level embeddings help capture both semantic and syntactic relationships. Additionally, optimization techniques like gradient clipping, regularization, and dimensionality reduction are applied to enhance model performance and efficiency.

1.5 Framework of the Research

The introduction presents an overview of BioNER, highlighting its significance, applications, challenges, and techniques.This section examines the methods and effects of transfer learning on biomedical research and healthcare and the Techniques section explores the deep learning methods employed in BioNER. The tools section details the various frameworks and tools used to implement BioNER systems. The challenges and issues section offers a thorough analysis of the hurdles in BioNER and possible solutions. The problem statement outlines the current limitations in BioNER systems, while the objectives section defines the aim of combining BioBERT and SciBERT to improve BioNER performance.

1.6 Existing Models in BioNER

Several models have been developed for BioNER. BioBERT, pre-trained on biomedical corpora, is highly effective in identifying biomedical entities. SciBERT, trained on scientific papers, excels in scientific contexts but is less specialized for biomedical tasks. BioWordVec utilizes Word2Vec and fastText for word embeddings, improving entity recognition through contextual word representations. Med7, built on SpaCy, efficiently processes clinical notes, while PubMedBERT, pre-trained on PubMed abstracts, is highly effective for biomedical text mining. The Biomedical Language Understanding Evaluation (BLUE) suite acts as a benchmark for assessing NLP models within the biomedical domain.. ClinicalBERT, fine-tuned on clinical notes, effectively extracts medical entities in healthcare settings. MT-BioNER, a multi-task learning model, enhances performance by sharing representations across tasks. CheXpert employs CNNs for radiology report analysis, specializing in radiology. MedT5, a transformer-based model, is used for text generation and classification in medical contexts. Combining LSTM with CRF improves sequence labeling in BioNER. Genia Tagger integrates rule-based and machine learning methods for entity extraction. The UMLS Metathesaurus is a knowledge-based resource for mapping biomedical concepts, while MetaMap maps biomedical texts to UMLS concepts. Lastly, BERT-CRF merges BERT's embeddings with CRF for sequence labeling, enhancing entity recognition accuracy.

1.7 Techniques

The proposed approach combines BioBERT and SciBERT through ensemble methods to capitalize on the specialized training of both models. A unified tokenizer and classifier from the Transformers library are used to streamline the process. Both models undergo fine-tuning for domain-specific tasks using techniques like multi-task learning (MTL) and conditional random fields (CRF) to enhance sequence labeling. Transfer learning is applied by leveraging pre-trained weights from both models, followed by specific fine-tuning on biomedical datasets to optimize performance.

1.8 Limitations

While this combined model enhances performance, its complexity may increase computational demands, though it is optimized for use on lower-tier GPUs. Integrating and deploying the model in real-world biomedical applications could present challenges, particularly in ensuring smooth interoperability between BioBERT and SciBERT.

1.9 Advantages and Disadvantages

The BioBERT and SciBERT ensemble model effectively captures long-distance dependencies in sequences, which is vital for tasks like Named Entity Recognition. It transforms words into dense numerical representations, improving the model's capacity to capture semantic and syntactic relationships, thereby enhancing its generalization abilities. By modeling dependencies between neighboring labels in sequence data, the model produces coherent predictions, improving accuracy in NER tasks. Additionally, parallel implementations boost training efficiency through concurrent processing, while sequential implementations may capture hierarchical task dependencies, leading to better generalization.

The model's complexity, however, can be resource-intensive, particularly when dealing with large datasets and complex architectures. Deep networks may face issues like vanishing or exploding gradients, and significant computational resources are required for training, particularly with large vocabularies or character-level embeddings. Pre-trained embeddings may not fully capture domain-specific nuances, necessitating additional fine-tuning. The model's complexity can also extend training and inference times and risk overfitting, especially with limited data or poorly selected features. Managing shared resources and synchronization in parallel implementations is crucial, while sequential implementations can introduce additional complexity and computational overhead.

II. RELATED WORKS

[1] S.-J. Yen et al, Introduced a machine learning-based framework for question answering to manage the growing volume of textual information available through modern information technologies and Internet services.

[2] M. Asghari et al, Incorporation of a model selector procedure with a hybrid indicator for online topic detection and an automatic data processing pipeline with regular and deep cleaning stages, utilizing multiple sources of meta- knowledge to enhance data quality.

[3] N. Vanetik et al, Proposed a novel concept of elementary discourse units (EDUs) selected based on parse tree gain, which improves the informativeness and conciseness of generated summaries.

[4] A. Passos et al, Presented a novel method for learning word embeddings that integrates information from

relevant lexicons, thereby enhancing representation quality.

[5] L. Ratinov and D. Roth. Explored, fundamental design challenges and misconceptions in developing efficient and robust NER systems, focusing on text chunk representation, inference methods, and the integration of prior knowledge sources.

[6] H. Fei et al, Introduced a dispatched attention neural model with multi-task learning to address the challenge of nested entities in named entity recognition.

[7] J. Devlin et al, Developed a novel language representation model pre-trained to capture deep bidirectional representations from unlabeled text, enabling state-of-the- art performance on various natural language processing tasks.

[8] J. Lee et al, Adapted a pre-trained language model for biomedical text mining, addressing challenges posed by unique word distributions in biomedical corpora.

[9] G. Crichton et al, Investigated the longitudinal effects of traumatic brain injury (TBI) on fatigue in children and adolescents, exploring the impact of time since injury and injury severity.

[10] X. Wang et al, Developed a framework for biomedical named entity recognition (BioNER) to overcome the limitations of current systems that rely on handcrafted features and limited training data for each entity type.

[11] S. Hong and J.-G. Lee, Introduced a novel CRF-based BioNER framework incorporating a deep learning-based label-label transition model to address the limitations of static representations.

[12] A. Marasović and A. Frank, Introduced AMTLHumor, an adversarial multi-task network utilizing neural network architectures based on Transformers to detect and rate humor and offensive texts.

[13] X. Ma and E. Hovy, Developed a novel neural network architecture that seamlessly integrates word- and character- level representations using bidirectional LSTM, CNN, and CRF, eliminating the need for hand-crafted features or data pre- processing in sequence labeling tasks.

[14] J.P. Chiu and E. Nichols, Discussed the quality and coverage of lexicons used in NER models and how the model's effectiveness can vary across different datasets or domains.

Despite advancements in Biomedical Named Entity Recognition (BioNER), current approaches often struggle with accurately identifying and classifying entities in complex biomedical literature. This study addresses this gap by proposing a novel solution: improving the precision and recall of BioNER systems through the integration of SciBERT and BioBERT, alongside the use of gazetteers to incorporate external domain knowledge. The key challenge lies in overcoming entity misclassification and incomplete recognition in biomedical texts. The proposed approach tackles these issues through an optimized model architecture and tokenization strategies, enhancing the robustness and accuracy of BioNER systems.

In conclusion, this research emphasizes the significant impact of advanced machine learning and natural language processing models on improving biomedical named entity recognition (BioNER). By integrating advanced models such as BioBERT and SciBERT with techniques like transfer learning and multi-task learning, the proposed methods improve the extraction and classification of crucial biomedical entities, including genes, proteins, diseases, and drugs.These approaches leverage robust embeddings and ensemble methods to enhance accuracy and generalizability across various biomedical subdomains. Despite significant advancements, challenges remain in terms of computational demands, model integration, and domain-specific nuances. Overall, these innovations promise to advance the field of biomedical text mining,leading to more effective and precise identification of key biomedical concepts, ultimately supporting better decision-making and research in healthcare and biomedical fields.

III. METHODOLOGY

The architecture of the proposed biomedical Named Entity Recognition (NER) system is designed to achieve high performance by integrating multiple components. It begins with a customized BERT architecture, optimized for token embeddings and tailored to handle the unique challenges of biomedical texts. BERT (Bidirectional Encoder Representations from Transformers) in the proposed framework is crucial for its ability to capture the contextual semantics of biomedical terms. Unlike traditional models, BERT is pretrained on a large corpus and fine-tuned specifically for tasks such as NER and relationship extraction. Its ability to process input bidirectionally means that it better understands the relationships between terms, even in complex biomedical literature, thus providing more accurate entity extraction.

Diverse datasets such as NCBI and BC5CDR are utilized to enrich the training process. The NCBI dataset contributes a wide range of biomedical literature, while the BC5CDR dataset, with its annotated texts on chemicals and diseases, adds depth and specialization to the training data.Reliability in biomedical knowledge extraction systems is paramount, especially given the critical nature of the domain. Ensuring reliability requires a multifaceted approach. First, using robust evaluation metrics across multiple datasets helps assess model performance comprehensively. These metrics, such as F1-score, precision, recall, and AUC-ROC, provide insights into the model's consistency across various data splits and domains, ensuring that it performs well not only in training but also in deployment scenarios. Additionally, incorporating explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations), allows researchers to interpret the model's decision-making process. This is particularly important in biomedical applications where trust in AI decisions is critical. These methods provide transparency, offering insights into why the model classifies certain tokens as entities, which in turn enhances the system's reliability.Another critical component is training on diverse, high-quality datasets from multiple biomedical domains. By curating datasets that reflect a wide range of biomedical subfields—such as disease annotations, chemical interactions, and protein-gene relationships—the model becomes more generalized, reducing the likelihood of overfitting to a particular dataset. This approach ensures that the system is robust and reliable, capable of handling diverse biomedical literature. Entity recognition within biomedical text is conducted through the use of Named Entity Recognition (NER), where the input text is tokenized and assigned labels following the BIO scheme. The BERT-based models (BioBERT, SciBERT) process the tokens to classify them as either Beginning (B), Inside (I), or Outside (O) of an entity. For instance, in identifying diseases, proteins, or chemicals, the model effectively learns to map contextual information around each token, leveraging the large corpus from which it was pretrained.This approach blends advanced machine learning techniques with the strategic integration of diverse datasets and ongoing performance monitoring, creating a robust system for biomedical NER tasks.

3.1 Performance Optimization Strategies

Improving performance in biomedical knowledge extraction can be achieved through several advanced techniques. The most effective approach is the use of transfer learning, particularly with domain-specific pretrained models like BioBERT and SciBERT. Fine-tuning these models with task-specific data, alongside hyperparameter tuning (such as adjusting learning rates and batch sizes), significantly enhances model performance. Additionally, using ensemble methods, such as combining outputs from multiple models, and incorporating gazetteer-based entity recognition further refines the results by reducing false positives and improving recall.

3.2 Evaluation metrics

For the evaluation, we utilized precision (P), recall (R), and F1- score (F1). Precision measures the model's ability to identify positive entities accurately. It is the ratio of correctly classified positive samples (True Positive) to the total number of classified positive samples. The higher the precision, the more accurate the prediction. Recall measures the model's ability to identify all positive instances correctly. This refers to the ratio of correctly pre- dicted positive samples to the total number of positive samples. The higher the recall, the more positive samples are detected. The F1-score represents the harmonic mean of precision and recall. Precision, recall, and F1-score are calculated using the following formulas:

$$P = TP / (TP + FP)$$
$$R = TP / (TP + FN)$$
$$F1 = 2 * (P * R) / (P + R)$$

3.3 Architecture

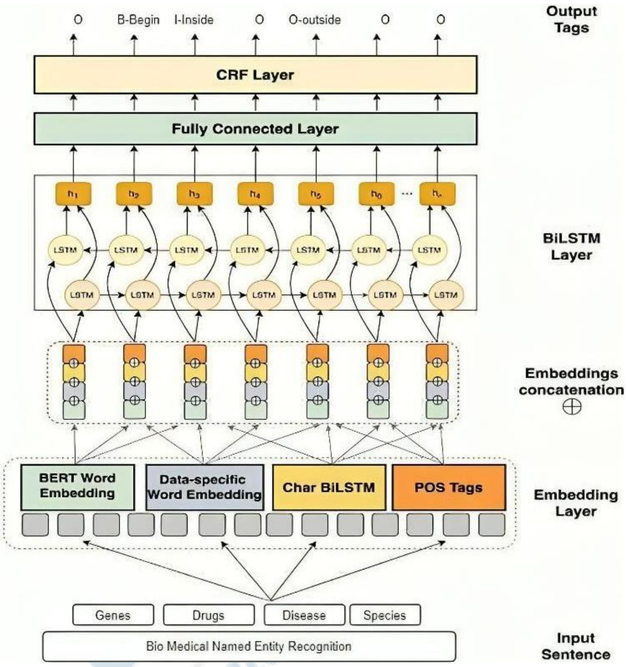


Figure 2. Architecture of the model
(Source: Hind Alamro, 2024)

The architecture outlined in the figure represents a sophisticated framework for Named Entity Recognition (NER) that integrates various embedding techniques and neural network layers to achieve high accuracy in entity tagging. The process begins with tokenizing the input sentence into individual words or subwords. These tokens are then processed through an embedding layer composed of several components: BERT word embeddings for rich contextual information, dataset-specific embeddings to refine representations, character-level BiLSTM embeddings for morphological features, and POS tags for syntactic context. These embeddings are concatenated to form a comprehensive representation for each token.

This concatenated representation is passed to a bidirectional LSTM (BiLSTM) layer, which captures dependencies from both past and future contexts, generating hidden states for each token. These hidden states are then passed through a fully connected layer, which reduces their dimensionality and prepares them for the final tagging stage.

A Conditional Random Field (CRF) layer is used at the output to ensure that the predicted tags follow valid transitions, thereby enhancing the overall accuracy of the sequence labeling.

The architecture effectively merges the strengths of pre-trained language models, dataset-specific embeddings, character-level features, and syntactic information, resulting in a robust NER system.

By utilizing CRF for sequential dependencies, it ensures precise tagging and significantly boosts the model's performance in recognizing entities within the input text. These embeddings include BERT word embeddings EBERT, data-specific word embeddings EData,character-level BiLSTM embeddings Echar and POS tag embeddings EPOS.

Econcat =EBERT + Edata + Echar + EPOS 3.1

This concatenated embedding Econcat is then fed into a bidirectional LSTM (BiLSTM) layer, which processes the sequence of embeddings. The BiLSTM captures both forward and backward contextual dependencies, producinghidden states ht for each token t.

ht = BiLSTM(Econcat,t) 3.2

where ht represents the hidden state at position t, capturing information from both previous and subsequent tokens in the sequence. These hidden states produced by the BiLSTM are then passed through a fully connected layer, transforming them into a format optimized for the final classification layer.

zt = wfc ht + bfc 3.3

where wfc and bfc represent the weight matrix and bias vector of the fully connected layer, respectively, and zt is the transformed hidden state. The output from the fully connected layer is subsequently fed into a Conditional Random Field (CRF) layer. The CRF layer models dependencies between the output labels, ensuring that the predicted tag sequence follows valid transitions. It computes the score for a sequence of tags y = (y1,y2,….,yT) given the input sequence of transformed hidden states Z=(z1,z2,…,zT).

S(y|z) = ∑Tt=1 WCRF (yt−1,yt) + ∑Tt=1 zt[yt] 3.4

where WCRF represents the transition scores between tags, and zt[yt] denotes the score of tag yt for token t.

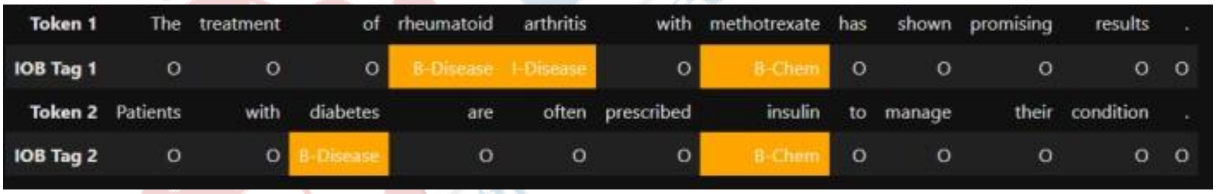
IV. RESULTS AND DISCUSSION

NCBI-Disease is a dataset fully annotated for diseases at both the mention and concept levels. The dataset includes 793 PubMed abstracts, 6,892 mentions of diseases, and 790 distinct disease concepts.

BC5CDR is a dataset created for the BioCreative V challenge. The dataset contains two sub-datasets: BC5CDR-Disease and BC5CDR-Chemical. We used these sub-datasets to evaluate diseases and chemicals, respectively.

Table 1. Number of sentences in Training, Validation, and Testing files in each dataset

Dataset	Entity type	Number of sentences			
		Training	Validation	Testing	Total
NCBI-disease	Disease	94,818	11,7391	124,676	230,585
BC5CDR-Chem	Chemical/Drug	135,615	23,959	24,488	184,062



Token 1	The	treatment	of	rheumatoid	arthritis	with	methotrexate	has	shown	promising	results	.
IOB Tag 1	O	O	O	B-Disease	I-Disease	O	B-Chem	O	O	O	O	O
Token 2	Patients	with	diabetes	are	often	prescribed	insulin	to	manage	their	condition	.
IOB Tag 2	O	O	B-Disease	O	O	O	B-Chem	O	O	O	O	O

Figure 3. Example of BIO tagging format

4.1 Improving Accuracy

To achieve high accuracy in biomedical NER, it is critical to implement techniques such as cross-validation, which helps ensure that the model generalizes well across different datasets. Moreover, expanding training datasets through data augmentation—generating synthetic examples or leveraging external biomedical datasets—can drastically improve the accuracy.

4.2 Improving Precision

Critical in reducing the number of false positives in biomedical NER tasks. One way to achieve this is by employing rule-based post-processing, where outputs from the model are filtered using predefined rules or constraints from biomedical dictionaries (gazetteers). This can be especially effective in ensuring that only valid entities are extracted. Moreover, adjusting the thresholds used in prediction can enhance precision without sacrificing recall. Preprocessing the data effectively, especially in complex biomedical contexts, ensures that the model captures nuanced terms and relationships, reducing error rates.

4.3 Improving Recall

Critical in reducing the number of false negatives in biomedical NER tasks. Improving recall ensures that all relevant biomedical entities are detected, which is essential in domains like healthcare, where missing key terms can lead to incomplete or inaccurate information extraction. One effective way to boost recall is by using comprehensive biomedical gazetteers during pre-processing, allowing the model to capture rare or less common entities. Additionally, lowering the prediction threshold can increase the model's sensitivity, improving recall at the potential cost of precision. Incorporating domain-specific knowledge and expanding the training dataset with diverse biomedical texts also helps the model generalize better, capturing more relevant entities.

Models are evaluated on diverse biomedical datasets, including NCBI and BC5CDR, which focus on chemicals and diseases. This evaluation ensures that the models are effective across various biomedical domains.

Table 5. Performance comparison between Biobert and Scibert Models

Metric/Report	SciBert			BioBert		
	P	R	F1	P	R	F1
B(Class)	0.8833	0.8784	0.8809	0.8880	0.8540	0.8706
I(Class)	0.7352	0.8084	0.7700	0.6876	0.7789	0.7304
O(Class)	0.9927	0.9916	0.9922	0.9912	0.9911	0.9912
Accuracy	-	-	0.9843	-	-	0.9824
Macro Avg	0.8704	0.8928	0.8810	0.8556	0.8747	0.8641
Weighted avg	0.9846	0.9843	0.9845	0.9828	0.9824	0.9826

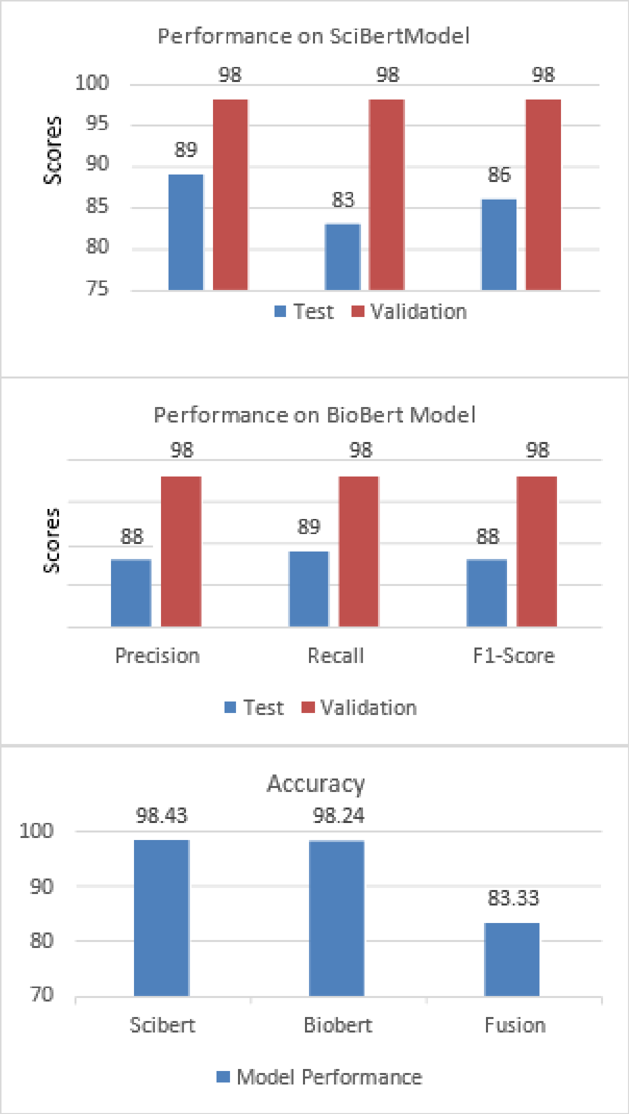


Figure 4. Performance comparison of using different BERT models with the (a) Scibert Model, (b) BioBert Model, and (c) Accuracy

The top-performing model is chosen for deployment, and post- deployment performance is monitored in real-time to track accuracy and other metrics, allowing for timely adjustments to maintain high performance.

Performance evaluation results show that SciBERT achieved a test accuracy of 98.43% and a validation accuracy of 98.52%. BioBERT had a test accuracy of 98.24% and a validation accuracy of 98.32%. The Fusion model, a novel technique introduced in this study, achieved a test accuracy of 83.33%, with validation accuracy not applicable. This analysis underscores the system's capability to leverage specialized biomedical datasets and advanced computational techniques for robust NER performance.

V. CONCLUSION AND FUTURESCOPE

The integration of SciBERT and BioBERT, a combination that leverages the unique strengths of each model, leading to improved entity recognition accuracy in biomedical texts. Second, we introduce an enhanced tokenization and alignment process that ensures better handling of complex biomedical terminology, resulting in a notable increase in performance metrics such as recall and F1-score. Additionally, the inclusion of gazetteers (external look-up lists) augments the model's ability to recognize domain-specific entities, contributing to improved precision. Lastly, a thorough ablation study was conducted to quantify the contribution of each model component, providing clarity on their individual impact on the overall system's performance.

5.1 Achieving Scalability

Scalability in the proposed framework can be achieved by leveraging modern GPU/TPU architectures for distributed training. As the size of biomedical datasets continues to grow, employing efficient parallel processing and model optimization techniques (such as mixed precision training) ensures that the model can handle large-scale data without compromising speed or performance. Transfer learning models, like BioBERT and SciBERT, also allow for rapid adaptation to new datasets with minimal additional training, enhancing scalability.

However, attempts to combine the strengths of both models have led to a decrease in accuracy, highlighting the need for more effective fusion strategies. Future efforts should aim to refine these techniques to more effectively combine the strengths of both models while minimizing performance declines. Advanced ensemble methods, such as stacking or weighted averaging, may provide more effective integration solutions. Additionally, refining hyperparameters, incorporating diverse and augmented datasets, and employing advanced evaluation metrics will be crucial for further enhancements. As computational resources advance, optimizing models for efficiency and scalability will increase their accessibility and application. Ongoing research and development in these areas are crucial for advancing NER technology and enhancing the accuracy and robustness of specialized text processing.

REFERENCES

[1] C.D. Santos, B. Zadrozny, Learning character-level representations for part-of-speech tagging, in: Proceedings of the 31st international conference on machine learning (ICML-14), 2014, pp. 1818–1826.

[2] H. Huang, H. Wang, D. Jin, A low-cost named entity recognition research based on active learning, Scientific Programming (2018).

[3] T. Mikolov, M. Karafiát, L. Burget, J. Cˇ ernocky´, S. Khudanpur, Recurrent neural network based languagemodel, in: Eleventh annual conference of the international speech communication association, 2010.

[4] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, arXiv preprint arXiv:1409.2329.

[5] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[7] N. Kitaev, D. Klein, Constituency parsing with a self-attentive encoder, arXiv preprint arXiv:1805.01052.

[8] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365.

[9] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[10] McCallum, D. Freitag, F.C. Pereira, Maximum entropy markov models for information extraction and segmentation., in: Icml, Vol. 17, 2000, pp. 591.

[11] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, J.-J. Liu, A support vector machine-based context-ranking model for question answering, Information Sciences 224 (2013) 77–8.